Stat 342 Exam 3
Fall 2014

I have neither given nor received unauthorized assistance on this exam.

KEY

_____

Name Signed **Date**

_____

Name Printed

There are 11 questions on the following 6 pages. Do as many of them as you can in the available time. I will score each question out of 10 points AND TOTAL YOUR BEST 7 SCORES. (That is, this is a 70 point exam.)

1. This is a "*no-intercept* one-variable linear regression" linear model problem. That is, suppose

$$y_i = \beta x_i + \varepsilon_i$$

for $i = 1, 2, \ldots, N$ where the ε_i are iid $\mathrm{N}(0, \sigma^2)$. $N = 5$ training cases are in the following table.

x	3	4	5	6	7
y	6	8.5	10	11.5	13

10 pts **a)** Find maximum likelihood estimates of the two (real-valued) parameters β and σ^2.

$X_{5\times 1} = \begin{pmatrix} 3 \\ 4 \\ 5 \\ 6 \\ 7 \end{pmatrix}$ $Y_{5\times 1} = \begin{pmatrix} 6 \\ 8.5 \\ 10 \\ 11.5 \\ 13 \end{pmatrix}$

$\hat{\beta}^{MLE} = (X'X)^{-1}X'Y$

$= (9+16+25+36+49)^{-1}$
$\times (18+34+50+69+91)$

$= \frac{1}{135}(262) = 1.9407$

$\hat{\sigma}^{2\,MLE} = \frac{1}{N}SSE = \frac{1}{5}\sum (y_i - \hat{y})^2 = \frac{1}{5}\left[(6 - 1.9407(3))^2 \right.$

$+ (8.5 - 1.9407(4))^2 + (10 - 1.9407(5))^2 + (11.5 - 1.9407(6))^2$
$\left. + (13 - 1.9407(7))^2 \right] \approx .2052$

10 pts **b)** Give 95% 2-sided prediction limits for an additional/future y based on a new value of the predictor, $x = 5.5$. (If you don't have an appropriate table of distribution percentage points with you, indicate *exactly* what % point of exactly what distribution (including d.f., etc.) you would put where in your calculation.)

How we need

$5.5(\hat{\beta}) \pm t\sqrt{MSE}\sqrt{1 + 5.5(X'X)^{-1}(5.5)}$

$\frac{1}{135}$

1.9407

upper 2.5% pt of t_{N-P} dsn i.e. $t_4 = 2.776$

$\frac{SSE}{N-P} = \frac{5(\hat{\sigma}^{2\,MLE})}{5-1} = 2.565$

i.e. 10.67 ± 4.92

2. One needs to do classification based on two discrete variables x_1 and x_2. A large training set (with $N = 100,000$) is summarized below in terms of **counts** of training cases with (y_i, x_{1i}, x_{2i}) of various types. (We implicitly assume these are based on an iid sample from the joint distribution of (y, x_1, x_2).)

$\hat{y}^{opt} = \arg\max_y g(y) f(\underset{\sim}{x}|y) = \arg\max_y f(y, \underset{\sim}{x})$ — here, these are estimated by cell counts over N

$y = 0$ cases

	$x_2 = 1$	$x_2 = 2$	$x_2 = 3$
$x_1 = 2$	9,000	3,000	10,000
$x_1 = 1$	6,000	4,000	8,000

$y = 1$ cases

	$x_2 = 1$	$x_2 = 2$	$x_2 = 3$
$x_1 = 2$	4,000	12,000	15,000
$x_1 = 1$	13,000	11,000	5,000

10 pts **a)** Of interest is an approximately optimal classifier based on the 100,000 training cases. Fill in the table below with either a 0 or 1 in each cell to indicate whether such a classifier classifies to $y = 0$ or to $y = 1$ for that (x_1, x_2) pair.

Simply compare the raw cell counts for $y=0$ to those for $y=1$ and pick the winner

Approximate $\hat{y}^{opt}(\mathbf{x})$:

	$x_2 = 1$	$x_2 = 2$	$x_2 = 3$
$x_1 = 2$	0	1	1
$x_1 = 1$	1	1	0

10 pts **b)** Compute a (0-1 loss) training error rate, $\overline{\text{err}}$, for your choice of classifier in **a)**.

This is just the number misclassified by the rule in a) divided by $N = 100,000$. This is

$$\overline{err} = \frac{1}{100,000}\left[4,000 + 3,000 + 10,000 + 6,000 + 4,000 + 5,000\right]$$

$$= .32$$

3. Here is a toy problem. Consider a 1-nn classification scheme based on a 1-dimensional input x. Suppose that a size $N = 5$ training set of pairs (y, x) is as below.

$$\mathcal{T} = \{(0,5),(1,6),(1,7),(0,7),(1,8)\}$$

Then suppose that 3 bootstrap samples are

$$\mathcal{T}^{*1} = \{(0,5),(1,6),(0,7),(1,8),(1,8)\}$$
$$\mathcal{T}^{*2} = \{(0,5),(0,5),(1,6),(1,7),(0,7)\}$$
$$\mathcal{T}^{*3} = \{(1,7),(1,7),(0,7),(1,8),(1,8)\}$$

10 pts **a)** Find the values of the 1-nn classifiers $\hat{y}^{*b}(x)$ based on each of the bootstrap samples for the training x's and record them below. (Break "ties" any way you wish.) Then give values for the bootstrap classifier $\hat{y}^{\text{boot}}(x)$.

	$x=5$	$x=6$	$x=7$	$x=8$
$\hat{y}^{*1}(x)$	0	1	0	1
$\hat{y}^{*2}(x)$	0	1	1 or 0	1 or 0
$\hat{y}^{*3}(x)$	1 or 0	1 or 0	1 or 0	1
$\hat{y}^{\text{boot}}(x)$	0	1	0 (or 1)	1

10 pts **b)** Find the OOB (out-of-bag) (0-1 loss) error rate for 1-nn classification based on this small number of bootstrap samples.

$\hat{y}^{*1}$ is built on all but (1,7) and makes 1 error on it

$\hat{y}^{*2}$ is built on all but (1,8) and makes, say, $\frac{1}{2}$ an error on that case

$\hat{y}^{*3}$ is built on all but (0,5) and (1,6) and makes, say, $2 \times (\frac{1}{2}) = 1$ error on them

Then an OOB error rate is

$$\frac{1}{4}\left[1 + \frac{1}{2} + \frac{1}{2} + \frac{1}{2}\right] = \frac{2.5}{4} = .625$$

numbers of errors made on missing cases

↑ total of numbers of cases missing in the y^{*b}

10 pts **4.** For the same toy scenario as in the previous problem, (1-nn classification based on the training set $\mathcal{T} = \{(0,5),(1,6),(1,7),(0,7),(1,8)\}$) find the 5-fold (0-1 loss) cross-validation error rate, CV.

Each individual data pair makes a "fold" and we thus need only find whether it's classified correctly by assigning to its nearest neighbor. So

$$CV = \frac{1}{K} = \frac{1}{5}\left(\frac{1}{1} + \frac{0 \text{ or } 1}{1} + \frac{1}{1} + \frac{1}{1} + \frac{0 \text{ or } 1}{1}\right)$$

(0,5) (1,6) (1,7) (0,7) (1,8)

which is, say, $= \frac{1}{5}\left(1 + \frac{2/3}{1} + 1 + 1 + \frac{1/2}{1}\right) = \frac{4\frac{1}{3}}{5} = \frac{13}{15}$

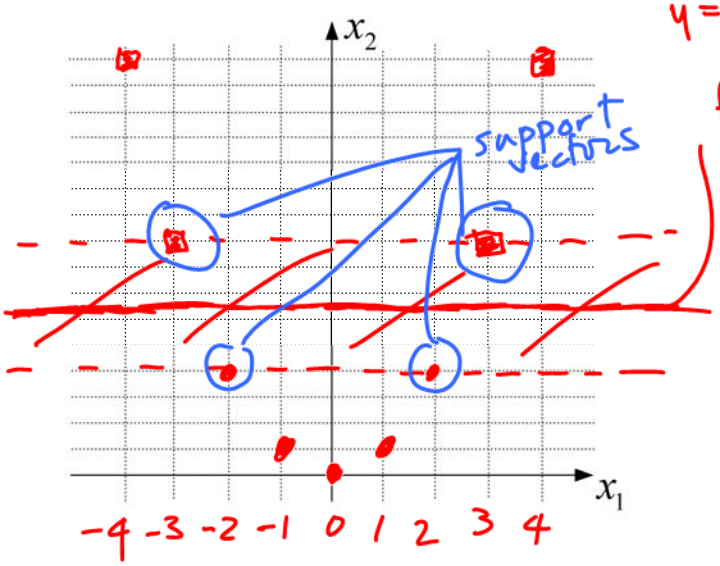
10 pts **5.** Below is another training set involving a 1-d predictor x. Argue carefully that there is no support vector classifier based only on x that has (0-1 loss) training error rate $\overline{\text{err}} = 0$. Then identify (by making a graph, not doing calculus) a maximum margin support vector classifier (with $\overline{\text{err}} = 0$) based on $x_1 = x$ and $x_2 = x^2$.

y	1	1	0	0	0	0	0	1	1
x	−4	−3	−2	−1	0	1	2	3	4

Rationale for no perfect SV classifier based on x alone:

A SV classifier based on x alone would split x at some value, say c, classifying 0 to one side and 1 to the other. For no c can one get all 1's on one side and all 0's on the other.

Maximum margin classifier based on $(x_1, x_2) = (x, x^2)$: $\hat{y}(x_1, x_2) = I[\beta_0 + \beta_1 x_1 + \beta_2 x_2 > 0]$



fat plane through the points centered at $x_2 = 6.5$ line

$\beta_0 = -6.5$ $\beta_1 = 0$ $\beta_2 = 1$

10 pts **6.** In a logistic regression problem with two standardized predictors x_1 and x_2, the model

$$\ln\left(\frac{P[y=0\,|\,\mathbf{x}]}{1-P[y=0\,|\,\mathbf{x}]}\right)=\beta_0+\beta_1x_1+\beta_2x_2$$

is fit twice, once by maximum likelihood and once by Lasso-penalized maximum likelihood. The two sets of fitted coefficients produced were

$\hat{\beta}_0=.44, \hat{\beta}_1=.98, \hat{\beta}_2=.66$ (run #1)

$\hat{\beta}_0=.52, \hat{\beta}_1=1.14, \hat{\beta}_2=.82$ (run #2)

Which of the two sets of coefficients do you think is the set of Lasso coefficients and why?

The first set has smaller $|\hat{\beta}_1|$ and $|\hat{\beta}_2|$ and therefore "less complex" fitted form for $\log\left(\frac{p}{1-p}\right)$. It is the Lasso version.

The fits were made using a data set with about the same numbers of $y=0$ and $y=1$ data points.

Based on run #1, what classifier would be appropriate for a case where it's known that $P[y=0]=.1$?

We go from a case-control data set to "the right" logistic regression form via $\hat{\beta}=\hat{\beta}^{cc}$ s.t. $\hat{\beta}_1$ and $\hat{\beta}_2$ don't change. We change $\hat{\beta}_0$ via

$$\hat{\beta}_0=\hat{\beta}_0^{cc}-\ln\left(\frac{N_0}{N_1}\right)+\ln\left(\frac{g(0)}{1-g(0)}\right)=.44-0+\ln\frac{1}{9}=-2.20$$

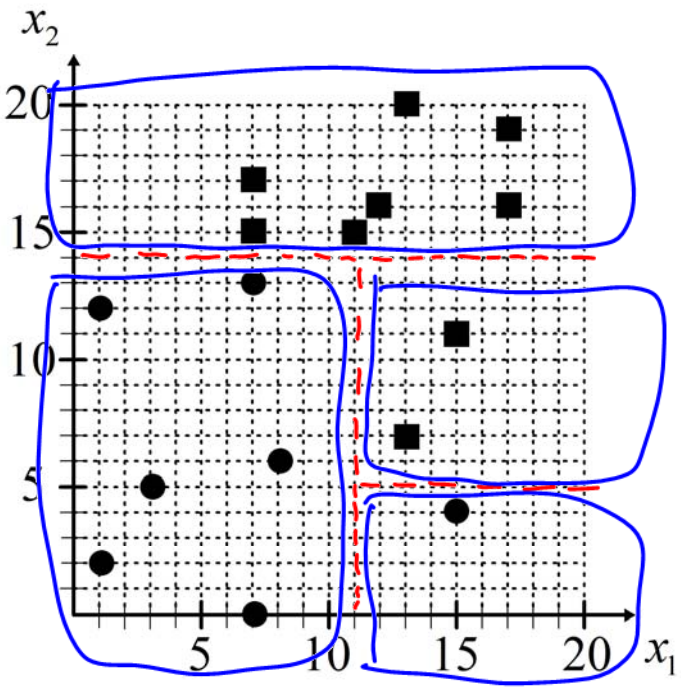
Then $\hat{y}(x)=I[-2.20+.98x_1+.66x_2<0]$

10 pts **7.** Suppose that 5 different classifiers are fit to a data set. If one wishes to make an ensemble classifier using weights $w_1=.3, w_2=.1, w_3=.2, w_4=.25,$ and $w_5=.15$ what "ensemble" classification is produced by each of the following sets of individual classifier results? (Write 0 or 1 in the empty cell for each set of $\hat{y}$'s. Write "0/1" for any "ties.")

$I[\text{weighted sum} > .5]$

.3 $\hat{y}_1$	.1 $\hat{y}_2$	.2 $\hat{y}_3$	.25 $\hat{y}_4$	.15 $\hat{y}_5$	$\hat{y}^{\text{Ensemble}}$
0	0	0	1	1	0
1	0	0	1	1	1
0	1	0	1	1	0/1
1	1	0	1	0	1
0	0	0	1	0	0
1	0	1	1	0	1
0	1	1	0	1	0
1	1	1	0	1	1
0	0	1	0	1	0
1	0	1	0	0	0/1

10 pts **8.** Below is a $p = 2$ classification training set for 2 classes.



y	x_1	x_2
0	1	2
0	1	12
0	3	5
0	7	0
0	7	13
0	8	6
0	15	4
1	7	15
1	7	17
1	11	15
1	12	16
1	13	7
1	13	20
1	15	11
1	17	16
1	17	19

Using empirical misclassification rate as your splitting criterion and forward selection, find a reasonably simple binary tree classifier that has empirical error rate 0. Carefully describe it below, using as many nodes as you need. Then **draw in the final set of rectangles** corresponding to your binary tree *on the graph above.*

At the root node: split on x_1 / x_2 (**circle the correct one of these**) at the value 14

Classify to Class 0 if $x_2 < 14$ (creating Node #1)

Classify to Class 1 otherwise (creating Node #2)

At Node # 1: split on x_1 / x_2 at the value 11

Classify to Class 0 if $x_2 < 11$ (creating Node #3)

Classify to Class 1 otherwise (creating Node #4)

At Node # 4: split on x_1 / x_2 at the value 5

Classify to Class 0 if $x_2 < 5$ (creating Node #5)

Classify to Class 1 otherwise (creating Node #6)

At Node # _____: split on x_1 / x_2 at the value _____

Classify to Class 0 if _____ (creating Node #7)

Classify to Class 1 otherwise (creating Node #8)

At Node # _____: split on x_1 / x_2 at the value _____

Classify to Class 0 if _____ (creating Node #7)

Classify to Class 1 otherwise (creating Node #8)

At Node # _____: split on x_1 / x_2 at the value _____

Classify to Class 0 if _____ (creating Node #9)

Classify to Class 1 otherwise (creating Node #10)